3rd URSI AT-AP-RASC, Gran Canaria, 29 May – 3 June 2022

Beyond labelled data: radio galaxy morphology classification using semi-supervised learning.

Inigo Val Slijepcevic*(1) and Anna Scaife(1)(2)
(1) Jodrell Bank Centre for Astrophysics, University of Manchester, Manchester, M13 9PL, UK
(2) The Alan Turing Institute, Euston Road, London, NW1 2DB, UK

1 Extended Abstract

Since initial publication over forty years ago, the Fanaroff-Riley classification scheme has continued to play an important part in the analyses of radio galaxies. In the past, classification of images into the FRI/FRII categories has largely been done *by eye*, with recent work leveraging various flavours of supervised convolutional neural networks (CNNs) as an effective replacement [1]. While this approach achieves high accuracy with enough labelled data, we do not have labelled data for new, large scale surveys. Producing a labelled catalogue for training is time intensive and any change in the labelling scheme requires a full re-label of the entire catalogue, incurring a high cost for experimenting with different classification schemes. In this work, we show that that by using semi-supervised learning, we can reduce the number of labels required for accurate radio galaxy classification. We attempt to use our approach “in the wild’ with unseen data from the same survey, but find that label imbalance in the unlabelled data degrades classifier accuracy.

We use FixMatch [2], a state-of-the-art algorithm designed to use unlabelled data alongside any available labels to train a custom CNN architecture optimised for radio galaxy classification. FixMatch relies on forcing consistency between its own predictions made on strongly augmented unlabelled images, allowing the model to learn from the data without labels. In our experiments we use data from the MiraBest dataset, a publicly available data-set with 1256 labelled radio galaxy sources [3], and the Radio Galaxy Zoo Data Release 1 (RGZ DR1) catalogue (Wong et al. in prep), which contains unlabelled sources.

Our results show that when using only the MiraBest data-set (for both unlabelled and labelled data), FixMatch is able to make accurate FRI/FRII classifications with just 51 labels, an improvement of almost an order of magnitude over the baseline supervised model. However, we find that using labelled data from MiraBest and unlabelled data from RGZ DR1 results in worse-than-baseline accuracy, which we show can be explained by label imbalance in RGZ DR1. This highlights an important obstacle when evaluating our models: training data has been chosen in a biased manner, which can manifest as a covariate or prior probability shift between the unseen data we wish to make predictions on, and the test we have used to evaluate model performance.

2 Acknowledgements

The authors gratefully acknowledge support from the UK Alan Turing Institute under grant reference EP/V030302/1. IVS gratefully acknowledges support from the Frankopan Foundation. This work has been made possible by the participation of more than 12,000 volunteers in the Radio Galaxy Zoo Project. The data in this paper are the result of the efforts of the Radio Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at `http://rgzauthors.galaxyzoo.org`.

References

[1] B. Becker, M. Vaccari, M. Prescott, and T. Grobler, “CNN architecture comparison for radio galaxy classification,” *Monthly Notices of the Royal Astronomical Society*, vol. 503, no. 2, pp. 1828–1846, 3 2021.

[2] K. Sohn, D. Berthelot, C. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in *Advances in Neural Information Processing Systems*, vol. 2020-Decem, 2020.

[3] F. A. M. Porter, “MiraBest Batched Dataset,” *https://zenodo.org/record/4288837*, 11 2020. [Online]. Available: https://zenodo.org/record/4288837